WHAT IS CLAIMED IS:

1. A server system comprising:

at least two scaleable tiers of server machines;

means for computing an average response time for the server system to respond to at least one transaction request; and

means for allocating a number of server machines for each tier of server machines such that the average response time for the at least one transaction request is less than or equal to a specified average response time.

2. The server system of claim 1 further comprising means for determining the costs associated with allocating the number of server machines at each tier of server machines.

3. The server system of claim 2 wherein said means for determining further comprises means for minimizing the costs associated with allocating the optimized number of server machines at each tier of server machines.

4. The server system of claim 3 wherein said means for minimizing comprises:

means operatively coupled to said server system for receiving input parameters and for solving:

$$\sqrt{\gamma} = \frac{\sum_{i=1}^{n} \sqrt{h_i s_i u_i}}{T - \sum_{i=1}^{n} s_i};$$

where: γ is the shadow price of the average response time; h_1, h_2, … h_n are weights reflecting the cost of different types of servers located at each tier of server machines; s is the average service time; u is the measured average utilization rate expressed in a single-machine percentage; and T is the average response time.

5. The server system of claim 1 further comprising at least one additional tier of server machines.

6. The server system of claim 1 further comprising:

a contractual relationship between a system operator and at least one contracting party; and

means for adjusting prices charged by said system operator to said at least one third party in response to a change in the allocation of server machines in at least two tiers of said server system.

7. The server system of claim 1 wherein said means for computing further comprises a non-iterative queuing model for predicting the average server system response time in response to measured arrival rates of transaction requests into each tier of server machines, the average service demand at each of said server tiers and the number of servers allocated to each tier of server machines.

8. A method for allocating a server machine to at least two tiers of a server system, said method comprising:

computing an expected average response time as a function of transaction requests and the amount of resources allocated to each tier of a server system;

determining whether an optimization problem is feasible;

computing a lower bound and an upper bound on the number of server machines at each tier of said server system required to meet the average response time; and

computing a solution specifying a number of server machines allocated to each tier of said server system such that transaction requests have an average response time less than or equal to a pre-determined limit.

9. The method of claim 8 wherein said computing an expected average response time further comprises:

obtaining at least one input value for an average arrival rate of transaction requests into each tier of said server system;

obtaining at least one input value for an average service demand at each tier of said server system; and

obtaining at least one input value for the number of server machines allocated at each tier of said server system.

10. A computer-readable medium comprising instructions for:

receiving selected input parameters representative of a server system having at least two tiers of server machines;

computing an average response time for the server system to respond to at least one transaction request; and

determining an allocation of server machines for each tier of server machines such that the average response time for the at least one transaction request is less than or equal to a specified average response time.

11. An assembly for allocating server machines in a server system comprising:

a pool of server machines;

at least two tiers of server machines;

means for computing an average response time for said tier of server machines to respond to a plurality of transaction requests; and

means for allocating a number of server machines from said pool to said tier of server machines to minimize operating costs while responding to said transaction requests with a specified average response time.

12. The assembly of claim 11 further comprising at least one additional tier of server machines.

13. The assembly of claim 11 further comprising:

a contractual relationship between a system operator and at least one contracting party; and

means for adjusting prices charged by said system operator to said at least one contracting party in response to a change in the allocation of server machines in said tiers of said server system.

14. The assembly of claim 11 wherein said means for computing further comprises a non-iterative queuing model for predicting the average server system response time in response to measured arrival rates of transaction requests into said tiers of server machines, the average service demand at said tiers of server machines; and the number of servers allocated to said tiers of server machines.

15. A server system comprising:

an open queuing network of multiple server machines with each server machine having a processor-sharing queue with a single critical resource;

at least two tiers of server machines; and

a computer-readable medium comprising instructions for:

(i) predicting the average system response time of said multiple server machines based on the arrival rate of transaction requests into each tier of server machines averaged over all transaction request types and the number of server machines allocated at each tier of server machines;

(ii)solving a mathematical representation of an optimization objective and constraints of said server system; and

(iii)determining a number of server machines for each tier of server machines in response to said predicted average system response time.

16. The server system of claim 15 wherein said mathematical representation comprises:

a continuous-relaxation model of the mathematical optimization system; and

an iterative bounding procedure.

17. The server system of claim 15 wherein said instructions for determining the number of server machines for each tier of server machines is in response to said predicted average system response time and at least one service level agreement (SLA) requirement.